Computer Vision Models Offer Scalable Species Detection From Social Media Photographs

Nathan Fox[1,2,*], Summer Mengarelli[3], Sabina Tomkins[4], Derek Van Berkel[2]

[1] Michigan Institute for Data & AI In Society, University of Michigan, USA
[2] School for Environment and Sustainability, University of Michigan, USA
[3] Hesburgh Library, University of Notre Dame, USA
[4] School of Information, University of Michigan, USA
[*] Corresponding author: foxnat@umich.edu

Abstract

Social media platforms have emerged as a promising source of data for biodiversity monitoring, due to the vast amounts of user-generated visual content. However, the unstructured and noisy nature of social media data poses challenges for accurate species identification. Foundation vision models present an innovative methodology for identifying a large diversity of species from photographs, however, they are yet to be robustly tested on messy social media data. This study explores the utility of foundation vision models in identifying species from social media images, focusing on charismatic species such as lions, cheetahs, and gorillas. We manually labeled a dataset of images from Flickr, taken in zoos across the United States, to establish a ground truth for species presence. We evaluated the performance of three models: (i) CLIP with binary prompts ("species name is present/species name is not present"), (ii) a categorical model with common object categories (e.g., “plant,” "building," "vehicle," and "expected species name"), and (iii) BioCLIP, a fine-tuned version of CLIP designed specifically for species identification. Our analysis revealed that the binary presence/absence model struggled with the noisy social media data, leading to low accuracy. The categorical model showed an improvement in true positive rates but continued to produce a large number of false positives. BioCLIP, while not achieving the highest accuracy, demonstrated superior performance in minimizing false positives, which is crucial for biodiversity monitoring where incorrect detections

can have significant consequences. Precision-recall analysis using presence-only data indicates their potential in real-world applications where presence detection is prioritized. Our findings suggest that foundation vision models show promise for scaling biodiversity monitoring through social media data.

Introduction

Biodiversity monitoring is crucial for understanding trends and drivers of biodiversity change, identifying effective conservation measures, and assessing progress toward global conservation targets, yet significant taxonomic, spatial, and temporal gaps hinder these efforts despite ongoing international commitments (Kühl et al., 2020). To increase our capacity to identify and record species sightings, there is a need for scalable approaches that overcome the limitations of traditional biodiversity surveying methods which are time-intensive even across smaller spatial-temporal scales (Schmeller et al., 2017). One such promising avenue is the growing interest in harnessing the vast amounts of visual data available through citizen-sourced image collections from social media websites, such as Flickr and Instagram (Ghermandi et al., 2023; Schirpke et al., 2023). Social media posts can be a useful indication of species sightings, adding to our understanding of their spatial-temporal distributions and patterns, including endangered, invasive, and migratory species (Allain, 2019; Barve, 2014; Fox et al., 2020; Hartmann et al., 2022; Jeawak et al., 2018; Sbragaglia et al., 2022). However, leveraging social media data for validating species sightings poses several challenges, particularly in terms of data quality and reliability.

Social media images, shared by users globally at high temporal resolution, present a unique opportunity to track species' presence across diverse geographic regions and timeframes. One of the primary challenges is the inherent messiness of social media data (Fox et al., 2020). Unlike citizen-sourced biodiversity datasets such as iNaturalist, where contributors often follow

specific protocols for identifying species to research-grade observations (Campbell et al., 2023; Di Cecco et al., 2021), social media posts are generally unstructured and lack validated species identification. Vague or incomplete descriptions may accompany images on platforms like Flickr, and the associated metadata, such as tags or titles, may not accurately reflect the content of the image (Barve, 2014; Fox et al., 2020).

Previous studies have often assumed that the presence of a species name in textual metadata is indicative of positive sightings; however, the lack of reliability in social media text identification means that without further validation, self-reported species identifications may be false positive sightings (Johnston et al., 2023). Often species validations of social media data rely on the manual inspection of images to confirm the presence of the intended species (Allain, 2019). However, manual validation is rarely feasible due to the vast size of social media datasets (Schirpke et al., 2021). The use of computer vision (CV), a branch of machine learning where computers can recognize and label objects in images, may allow for rapid and automatic identification of species from within images (August et al., 2020).

CV models can accurately detect species from photographs, across the diverse taxa of flora to fauna (Wäldchen & Mäder, 2018; Weinstein, 2018). Many of the current CV models for species identification are supervised learning models, which are built using labeled photographs of species to train the model (Weinstein, 2018). However, building accurate CV models requires a large labeled dataset of example images for each species, which may not be readily available (Fernandes et al., 2020). Furthermore, CV models are often fine-tuned to identify a single species, several species, or specific geographic contexts (Weinstein, 2018). With the vast amount of species across diverse geographic contexts captured in social media imagery, it is

unlikely that the application of supervised learning models will be transformative in providing rapid validation of species from social media.

Given these challenges, the use of foundation vision models may increase our capacity for species identification from social media images. Foundation models are those in which a model is trained on diverse data, allowing it to be applied across a wide range of use cases (Li et al., 2024). For CV tasks, the innovative CLIP (Contrastive Language-Image Pretraining) model is a machine-learning model developed by OpenAI that can understand images and text together (Radford et al., 2021). These models can generalize across various tasks, including species identification, without requiring task-specific fine-tuning. Fine-tuning is a process where a pre-trained model is further trained on a specific, often smaller, dataset (i.e., images of tigers) to enhance it for more specialized tasks while retaining the general knowledge acquired during the initial training. The key feature of CLIP is its ability to work in a zero-shot manner, meaning it can make predictions on tasks it was not explicitly trained for by leveraging its understanding of language. For example, it can classify images based on text prompts like "a photo of a tiger" without having been trained on a specific dataset of tiger photos. In benchmarking tests, the CLIP model outperformed other CV models in identifying certain taxa such as birds and flowers (Radford et al., 2021).

Though CLIP is a robust tool for identifying certain species, its general training and benchmarking, which were aimed at a variety of tasks all within specific contexts, may not generalize effectively to the identification of other species (Radford et al., 2021). To improve the potential shortcomings of CLIP for species identification, the BioCLIP model was trained specifically for species identification (Stevens et al., 2024). Leveraging the same underlying architecture as CLIP, BioCLIP has been trained on a vast dataset that includes over 450,000 species, enabling it to accurately recognize and classify various plants, animals, and other

organisms from images. This specialized training makes BioCLIP particularly effective for tasks involving identifying species in complex, real-world scenarios that might extend to social media images. By combining the flexibility of CLIP with domain-specific knowledge, BioCLIP offers a powerful tool for ecological research, conservation efforts, and biodiversity data analysis. BioCLIP's benchmarking tests saw significant improvements in species identifications compared to other CV models (Stevens et al., 2024).

Though both CLIP and BioCLIP have previously demonstrated robust methods for species identification, the effectiveness of these models in handling the unique challenges posed by social media data has yet to be tested and remains an open question. This study, therefore, aims to assess the utility of foundation vision models for species identification in social media images. By comparing the performance of three different models: CLIP with binary prompts, CLIP with categorical prompts, and BioCLIP, we seek to determine which approach offers the best balance of accuracy and reliability for identifying species from social media images.

Methods

Dataset Collection

For this study, we selected a dataset of images from Flickr, a popular social media platform where users frequently share photos of wildlife. To ensure a focused analysis, we limited our selection to images taken in zoos across the United States (Fig. 1). We used the photosearcher R package (Fox et al., 2020) to return all Flickr images from within the footprints of zoos within the US taken from OpenStreetMap (OpenStreetMap, 2024). We then extracted any of the images that contained taxa common names within the photographs' titles, descriptions, or tags. Our final list of assessed species included two big cats (*lion and cheetah*), two primates (*gorilla,*

orangutan), three birds (*flamingo, ostrich, penguin*), one bear (*polar bear*), and one marsupial (*koala*).

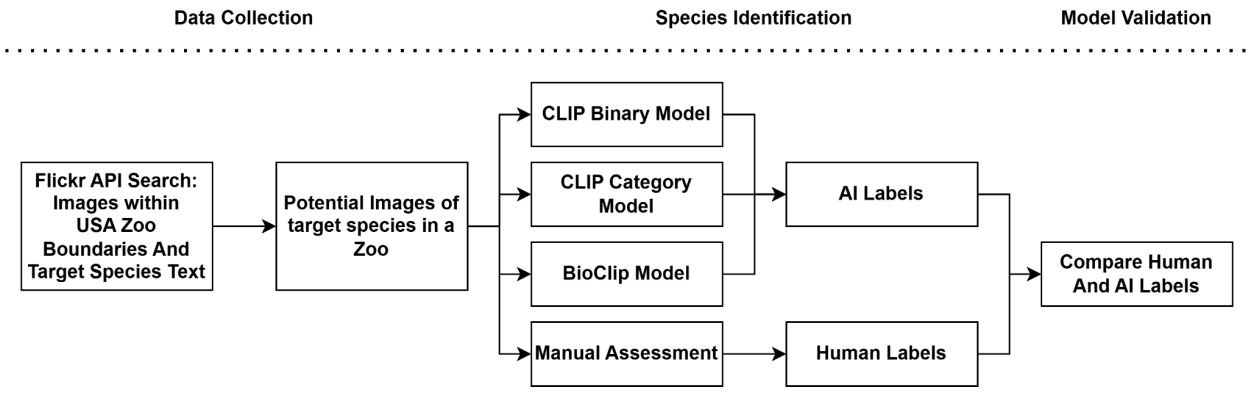

Figure 1. Overview of methods: data collection from Flickr API using a zoo boundary shapefile and filtering results by a target species name, manual and computer vision model labeling of species presence in photographs, and comparing human and AI labels.

We opted to use images taken in zoos as our primary dataset due to the unique combination of expected and challenging conditions they present. Zoos are environments where the presence of certain species is highly likely, making them a controlled setting for assessing model performance. However, zoo images also tend to be messy, reflecting real-world challenges such as incorrect species tagging, the inclusion of multiple species in a single image, images of people, features of their captivity that may obscure them (e.g., behind cages or glass), and the potential for images of non-biological subjects (e.g., zoo signage, buildings) to be mislabeled with animal names (Fox et al., 2020; Kulkarni & Di Minin, 2023; Spooner & Stride, 2021). This mix of expected species presence and inherent data noise provides a robust test for the model's ability to accurately identify species in unstructured social media data.

The species selected for this study were chosen to represent a diverse range of taxa, thereby allowing us to assess the generalizability of the foundation vision models. This cross-taxa

selection is crucial for evaluating whether the models can reliably identify species beyond a narrow taxonomic focus. Furthermore, due to the biases in what species citizens choose to upload to Flickr (Marshall & Strine, 2019), these species were chosen due to their widespread recognition and the likelihood of being featured in both images and associated text (Mangachena & Pickering, 2023; Tenkanen et al., 2017). This provided a reliable test-bed of accurate species tags, which are often lacking in social media content, for testing these foundational models.

Manual Labeling Process

To establish a ground truth for the presence or absence of the targeted species in each image, we manually labeled the dataset. For each of the selected taxa, if the species name was mentioned in the textual metadata, we visually assessed the photograph for that taxa. We then labeled the images as present or absent for that taxa. Most taxa were identified to the species level; however, we note that where the searched taxa do not belong to one specific species (e.g. gorilla, flamingo, and penguin), here were only manually identified to the higher taxonomic level (e.g. family, or genus).

To establish a reliable ground truth for the presence or absence of the target species in the images, we employed a two-step validation process. In the first step, one author conducted the initial labeling of each image. In the second step, a second author verified these labels independently. If discrepancies were identified during the verification process, the image would be re-evaluated jointly to reach a consensus. However, in this study, there was complete agreement between the two authors, with no discrepancies found. This approach was complemented by the use of predefined criteria for species identification, focusing on distinctive morphological characteristics and contextual cues. Although inter-rater reliability metrics like Cohen’s kappa were not applicable due to the nature of the process, the two-step confirmation

method ensured a robust and accurate labeled dataset, which served as the ground truth for evaluating the computer vision models.

If there was no visual identification of the taxa, the image was labeled as not containing that species. This manual labeling process was critical for evaluating the accuracy of the foundation vision models used in the study. Whilst manually labeling the images, we also noted frequent objects in images where a species was mentioned but not present (e.g. a photo of a building). We used these false sightings to inform the categorical model's candidate labels.

Foundation Vision Models

We evaluated three foundation vision models for their effectiveness in identifying species from the labeled images. Specifically, we use the clip-vit-large-patch14 zero-shot-image-classification mode through the 'transformers' Python library (Wolf et al., 2020), and BioCLIP TreeOfLifeClassifier through the pybioclip Python library (Bradley et al., 2024). We used the following methodologies to test the foundation vision model's ability to identify species:

1. **CLIP with Binary Prompts:** The first model we used was CLIP (Contrastive Language–Image Pretraining), a versatile vision-language model. For each image, we applied binary prompts, such as *"lion is present*" and *"lion is not present,"* to determine the likelihood of the target species being present in the image. This method leveraged CLIP's ability to match textual descriptions with visual content.
2. **Categorical Model:** The second model applied a set of predefined categories, including, *"animal," "plant," "landscape," "human," "vehicle," "building," "object," "food," "drink," "art,"* and then the specific species, e.g. *"lion."* These labels were informed from the manual labeling process in which the authors noted objects that often appeared in misstaged photographs. This model was designed to identify whether an image could be

categorized as containing a specific object, such as buildings, a generic plant or animal that was not the target species, or the specific species we were after. This approach allowed us to test the model's ability to differentiate between broader categories and our specific species.

3. **BioCLIP:** The third model, BioCLIP, is a fine-tuned version of CLIP, specifically adapted for species identification. As BioCLIP can only identify taxa, and not additional categories, such as buildings, we employed its open-ended model to return the most likely species present in the image.

Evaluation Metric

To compare the performance of the three models, we calculated true positives, false positives, true negatives, and false negatives. True positives represented instances where the model correctly identified the presence of the species, while false positives indicated cases where the model incorrectly identified the species as present. Conversely, true negatives and false negatives reflected the model's accuracy in identifying images where the species was absent. These metrics provided a comprehensive overview of each model's strengths and weaknesses in handling unstructured data from social media. When evaluating the model accuracy, we only expected the AI to match the same taxonomic rank as the human label. For example, with CLIP, we prompted it to label “flamingo” without specifying a subspecies such as the lesser flamingo, while with BioCLIP, the identification of any of the six recognized flamingo species was accepted as a match. To visualize model agreement and disagreement between the best two models, we aggregated prediction results for nine species using two models: CLIP Categorical

and BioCLIP. Binary predictions were compared to human-labeled presence/absence to calculate agreement and disagreement across all species.

Precision-Recall Analysis Using Presence Data Only

To focus on the models' ability to detect the presence of species in noisy, real-world social media images, we calculated the precision-recall using the subset of data labeled as 'present' (i.e., cases where a species was actually present in the image). This was done to better assess the models' performance in detecting species presence, aligning with real-world conservation tasks where the detection of presence is critical for monitoring endangered or invasive species. By limiting the analysis to presence data, we minimized the influence of absent cases that could otherwise inflate precision and skew the interpretation of model performance.

For each model, we extracted the confidence scores associated with the species presence predictions. Precision-recall curves were generated by plotting precision (the proportion of true positive predictions among all positive predictions) against recall (the proportion of true positives among all actual positives) at different confidence thresholds. The Average Precision (AP) score was calculated as a summary metric, representing the area under the precision-recall curve.

Results

In total, we collected and labeled a dataset comprising 13,230 images from Flickr. Among these, the lion was the most common species, with 6,150 images, followed by flamingos with 1,589 images, and penguins with 1,113 images. The CLIP Binary model showed an average accuracy of 52.96% with a standard deviation of 14.90%, indicating higher variability in performance

across species (Table. 1). The Clip Categorical model, on the other hand, achieved the highest average accuracy of 91.13% with a lower standard deviation of 4.18%, reflecting consistently high performance with minimal variability. BioCLIP demonstrated a moderate average accuracy of 74.70% and a standard deviation of 10.32%, suggesting a balanced but variable performance across different species.

Table 1. Overall Model Accuracies Per Species

Species	**CLIP Binary**	**Clip Categorical**	**BioCLIP**
Cheetah	52.8%	96.4%	79.73%
Flamingo	35.05%	92.01%	85.08%
Gorilla	55.23%	93.60%	56.20%
Koala	53.30%	87.74%	81.13%
Lion	42.70%	93.54%	89.22%
Orangutan	64.00%	96.55%	61.38%
Ostrich	67.45%	87.26%	75.94%
Penguin	28.21%	89.67%	67.65%
Polar bear	77.92%	83.44%	75.97%

The CLIP Categorical model is the best performer in terms of overall accuracy, particularly excelling in species like cheetahs and lions. However, this comes at the cost of a higher number of false positives (Figure 2), which can introduce noise into datasets when used for large-scale biodiversity monitoring. On the other hand, BioCLIP, while not achieving the highest overall

255 accuracy, excels in minimizing false positives, making it particularly valuable when precision in
256 species identification is critical. The binary model consistently underperformed relative to the

other two models, highlighting its limitations in dealing with the complexities of social media images.

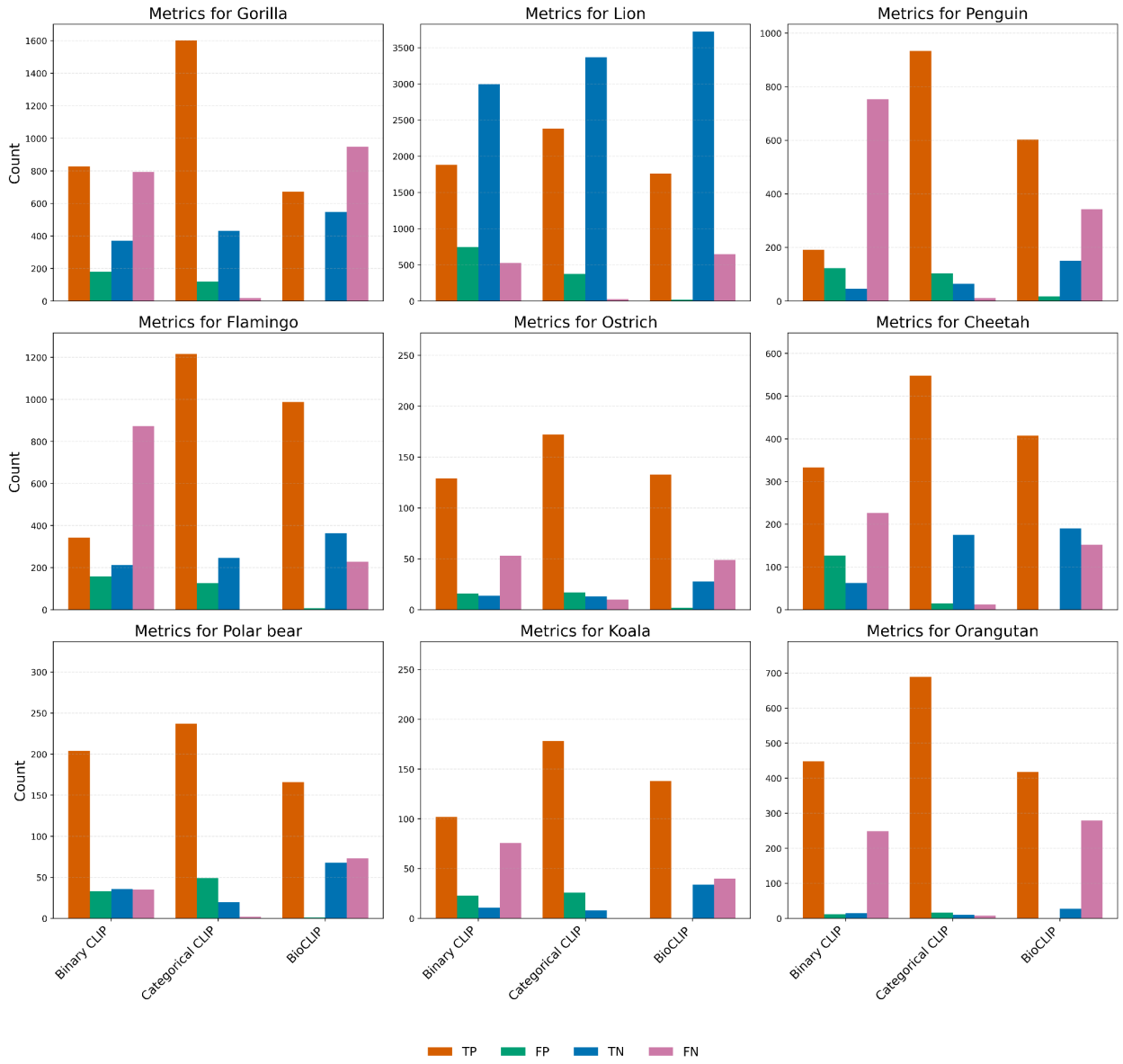

Figure 2. True positives, false positives, true negatives, and false negatives of the three model types (CLIP with binary prompts, CLIP with categorical prompts, and BioCLIP) across each species.

The CLIP Binary model, which utilized simple presence/absence prompts, exhibited significant variability in performance across different species. The model's highest accuracy was observed

in the ostrich category at 67.5%, where it correctly identified 129 true positives and maintained relatively low false positive counts (16 FP). However, the model struggled considerably with species like penguins, where it achieved an accuracy of only 28.2%, marked by a large number of false negatives (753 FN) and false positives (46 FP). Overall, the binary model showed limited effectiveness, particularly in handling the noisy and diverse nature of social media data, resulting in generally lower accuracy across most species.

The categorical model, which categorized images into broader groups like "animal" and specific species names, emerged as the most accurate overall. This model achieved the highest accuracy for species like cheetah, with an impressive 96.4% accuracy, supported by 548 true positives and a relatively low count of false negatives (12 FN). Similarly, lion identification also performed well, with an accuracy of 93.5%, indicating strong reliability in identifying this species across a large dataset. However, despite its higher overall accuracy, the model still exhibited a higher number of false positives, particularly in more ambiguous categories. For example, penguin identification, while achieving 89.7% accuracy, still suffered from 104 false negatives, indicating that the model's broader categorizations sometimes led to overgeneralization.

BioCLIP, the fine-tuned model specifically designed for species identification, showed a balanced but slightly lower overall accuracy compared to the Clip Categorical model, with its best performance in minimizing false positives. For example, flamingo identification yielded an accuracy of 85.1%, with only 8 false positives—significantly fewer than those observed in the categorical model. However, BioCLIP's overall accuracy for species like lions was 89.2%, slightly lower than the categorical model, but it had a much lower false positive count (17 FP), highlighting its strength in reducing incorrect identifications. Gorilla identification showed one of

the lower performances for BioCLIP, with an accuracy of 56.2%, yet it still outperformed the binary model by maintaining a low number of false positives.

Model Agreement

The aggregated results between BioCLIP and the CLIP Categorical model revealed strong agreement (Fig. 3), with 5289 true positives (both models predicting "Present") and 4337 true negatives (both predicting "Absent"). However, disagreements were observed primarily in false negatives, where BioCLIP predicted: "Absent" while CLIP Categorical predicted "Present" (2670 cases). False positives were less frequent, with 797 instances of "Present" being predicted by CLIP Categorical while labeled "Absent" by human annotations. The models showed minimal disagreement in negative predictions, with zero cases where BioCLIP predicted "Present" while

CLIP Categorical predicted "Absent." These results highlight the overall consistency between the models while emphasizing the need to improve sensitivity for detecting true positives.

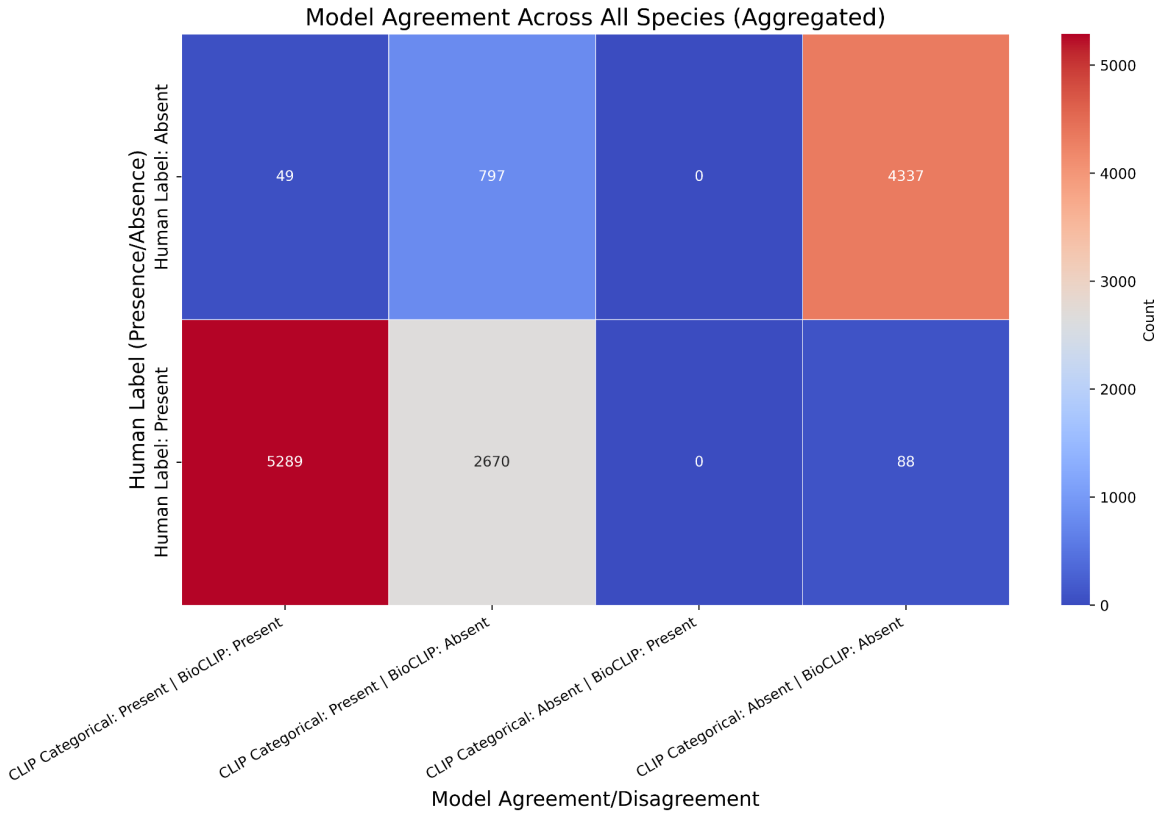

Figure 3. Aggregated heatmap showing model agreement and disagreement between CLIP Categorical and BioCLIP predictions across all species. Rows represent human-labeled presence or absence, while columns show agreement/disagreement categories for the models. Counts in each cell indicate the number of predictions for each combination, highlighting strong agreement in true positives and true negatives and discrepancies in false negatives and false positives.

Precision-Recall Analysis Using Presence Data Only

The evaluation of the three models—CLIP Binary, CLIP Categorical, and BioCLIP—across nine species revealed notable performance differences in handling presence detection using social

media data (Fig. 4). Overall, the CLIP Categorical model consistently achieved the highest average precision (AP) scores, excelling with species like lion (AP = 0.82), penguin (AP = 0.92), and Cheetah (AP = 0.99). Its broader categorical classification approach proved highly effective for general species detection tasks. The BioCLIP model, while specifically designed for species-level identification, showed strong performance in species like ostrich (AP = 0.93) and flamingo (AP = 0.91), yet demonstrated variability in precision across species due to inconsistent confidence calibration at higher recall levels. In contrast, the CLIP Binary model struggled

overall, particularly with more challenging species like gorilla (AP = 0.67) and Lion (AP = 0.35), highlighting its limitations in binary presence detection.

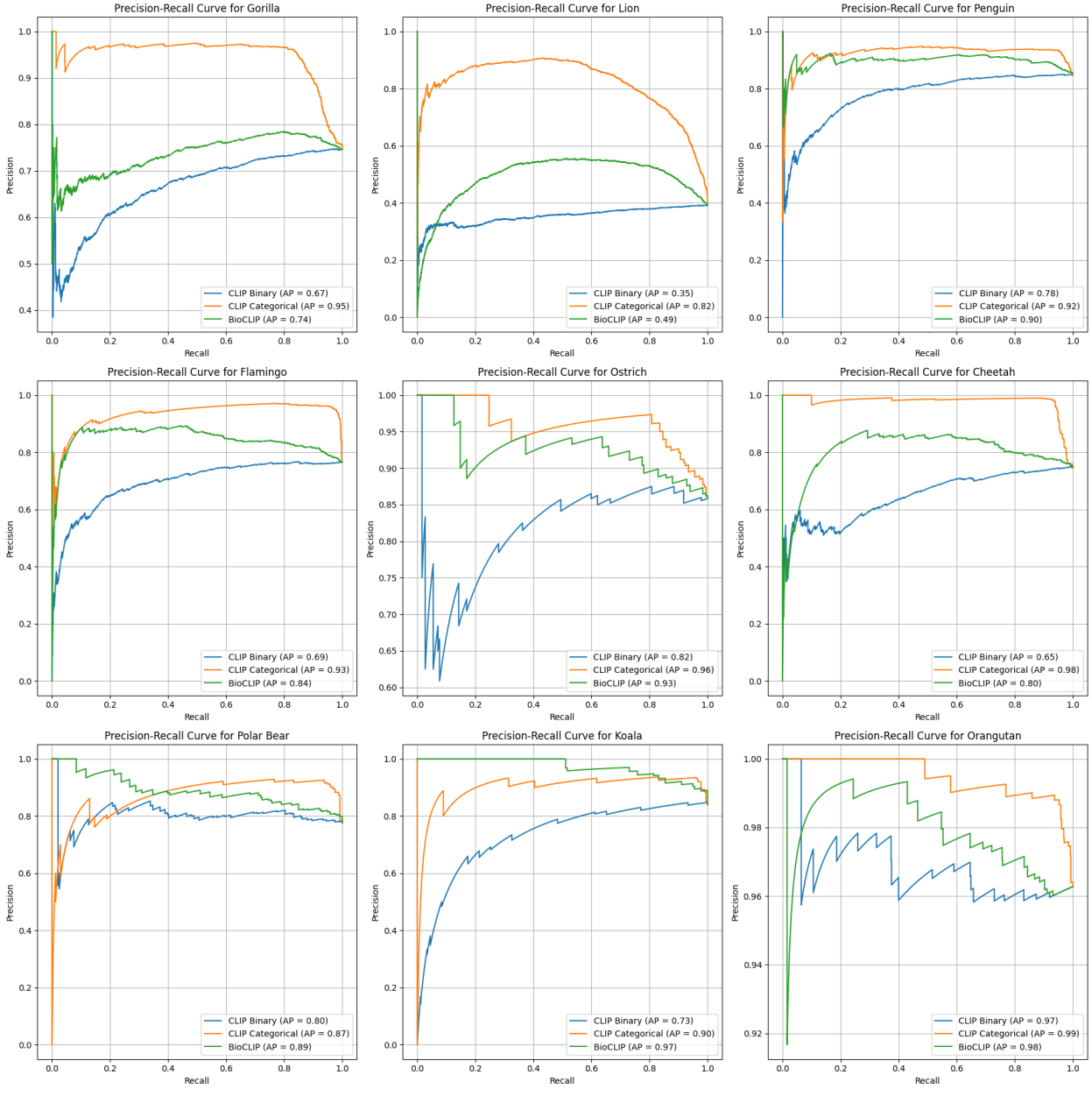

Figure 4. Precision-Recall Analysis Using Presence Data Only

Interestingly, while the precision-recall curves indicated strong performance for species like orangutan, where both CLIP Categorical and BioCLIP models achieved near-perfect curves (AP = 0.99), this did not always translate into high overall accuracy scores. For example, despite the

near-perfect precision-recall curves for identifying orangutans, the CLIP Binary model had an overall accuracy of only 64%. This discrepancy arises from the difference between the metrics: precision-recall curves focus on evaluating a model's ability to correctly identify positive cases (i.e., species presence), which is especially useful in datasets where absent cases dominate. In contrast, accuracy reflects the proportion of all correct predictions (both presence and absence), making it more sensitive to errors in detecting absent cases.

Discussion

Foundation vision models such as CLIP and BioCLIP provide scalable solutions for significantly enhancing species identification from large social media image datasets. The relatively high accuracy, coupled with fast run times, offers a substantial advantage over traditional methods that are often labor-intensive and geographically constrained (Ghermandi et al., 2023; Schirpke et al., 2023). Data from social media may therefore be particularly valuable in expanding the coverage of species monitoring efforts to under-represented regions, potentially filling critical gaps in data availability (Soriano-Redondo et al., 2024; Toivonen et al., 2019). These validated sightings from social media can then be used for more nuanced biodiversity assessments including shifting ranges due to climate change, monitoring migratory patterns, assessing the spread of invasive species, tracking illegal wildlife trade, and understanding public sentiment towards wildlife (Allain, 2019; Cardoso et al., 2024; Mancini et al., 2019; Sbragaglia et al., 2022).

Our comparative analysis of the three models (CLIP with binary prompts, CLIP with categorical prompts, and BioCLIP) reveals important insights into their relative strengths and weaknesses. Though it may be expected that foundation models, such as CLIP, may be better suited to simple binary classifications (Shen et al., 2021), our binary presence/absence model was the least accurate in species identification. That said, the model was accurate for some species,

demonstrating that a simple binary classification may be useful in some species or environmental contexts. By designing categorical prompts based on other expected objects, our categorical model showed improved performance compared to the binary model. With the known context of likely objects in the photograph set, effective prompt designs can help to improve the model accuracy (Wang et al., 2023). Our findings suggest that the categorical model is generally the best choice for maximizing true positives and overall accuracy.

BioCLIP, which is specifically fine-tuned for species identification, demonstrated the highest reliability in minimizing false positives. This is of particular importance given that false-positive observations in biodiversity monitoring can be more consequential than false negatives (Groom & Whild, 2017). Although BioCLIP did not achieve the highest overall accuracy, its accuracy in reducing incorrect identifications suggests that it may be the most suitable option for automatically generating "validated" datasets. It is important to note that BioCLIP's focus on species identification means it may still generate false positives in images that do not contain biological subjects, such as buildings or vehicles. This limitation underscores the need for robust filtering and validation of social media data for ecological studies (Fox et al., 2021). BioCLIP's strength in minimizing false positives makes it a valuable tool in contexts where accuracy is prioritized over broad coverage. The choice of model should therefore be guided by the specific requirements of the study, whether the emphasis is on maximizing correct identifications or minimizing false positives.

The models' high precision-recall scores for species indicate that they are quite effective at identifying the species when present. However, the overall accuracy is affected by misclassifications in "absent" cases. For instance, a model might perform well in identifying "present" cases but struggle with predicting the correct "absent" labels, leading to lower accuracy despite high AP scores. This discrepancy is especially apparent in datasets where a

substantial imbalance exists, with many more "absent" labels than "present" ones. Thus, while precision-recall curves provide a nuanced view of species detection capabilities, especially for conservation-focused tasks where presence detection is prioritized, accuracy scores reveal potential pitfalls in overall classification robustness.

Using both the CLIP Categorical and BioCLIP models in tandem enables the generation of highly accurate presence and absence datasets where their predictions align, reducing the need for extensive manual validation. Disagreement between the models highlights areas that require human intervention, ensuring that critical errors are addressed efficiently. This dual-model approach strikes a balance between automation and precision, minimizing the manual effort required while maximizing the reliability of the final dataset. By leveraging model agreement for validation and targeted human input for resolving discrepancies, this method offers a scalable solution for large biodiversity datasets.

Despite the promise of these models, it is also important to acknowledge the limitations and biases within the data and models, particularly regarding the quality and reliability of social media data. Though here we attempted to capture a range of taxa, due to the biases in social media uploads representing a small number of taxa (Edwards et al., 2021), our photographs from zoos mainly captured charismatic mammals and bird species. Furthermore, these models may not accurately identify the diversity of species in the range of image contexts encountered in real-world applications. For instance, though BioCLIP can identify over 450,000 species (Stevens et al., 2024), this does not fully capture the vast number of species on Earth (Wiens, 2023). Future efforts should, therefore, enhance our understanding of the accuracy of using

foundation vision models to label harder-to-identify species, such as rare or similar-looking species.

Given these limitations, foundation vision models must be viewed as complementary tools rather than replacements for traditional manual identification and expert-verified datasets (Toivonen et al., 2019). The integration of these models with traditional data sources can help mitigate biases and enhance the overall quality and utility of biodiversity data. While the scalability and global reach of these models are clear, their true value lies in their ability to complement, rather than replace, existing biodiversity monitoring systems (Soriano-Redondo et al., 2024). Traditional datasets, including those derived from citizen science platforms like iNaturalist, provide structured, high-quality data that is critical for accurate species identification. By integrating social media and foundation vision models with these traditional datasets, researchers can enhance the breadth and depth of biodiversity data, leading to more comprehensive and informed conservation strategies.

One of the most promising opportunities is the ability to capture a broader range of species observations at a much finer temporal and spatial resolution than traditional methods alone can achieve (Fox et al., 2024). Social media platforms, with their vast user base, generate an ongoing stream of biodiversity observations that are often geotagged and timestamped. This continuous stream of data can fill crucial gaps in traditional datasets, especially in regions where formal surveys are scarce or where resources for fieldwork are limited. By leveraging foundation vision models to process this influx of unstructured social media data, researchers can rapidly

identify new species occurrences and compare them with historical ecological datasets to detect shifts in species distributions and patterns (Fox & Van Berkel, 2024).

Moreover, the integration of these diverse data sources enables a more holistic understanding of ecosystems, extending beyond species presence to include behavioral insights that might not be captured in conventional surveys. For example, social media posts can reveal nuanced patterns of species interactions, such as predator-prey dynamics or seasonal changes (August et al., 2020; Tuia et al., 2022). By aligning these insights with the structured, high-quality data from traditional surveys, researchers can develop a more nuanced, multi-layered approach to conservation planning. Ultimately, this synergy not only accelerates biodiversity assessments but also strengthens the scientific foundation for adaptive management strategies that are essential in the face of rapid environmental changes (Fox et al., 2024; Toivonen et al., 2019).

Conclusions

Foundation vision models such as CLIP and BioCLIP offer a promising avenue for advancing biodiversity monitoring, particularly when integrated with traditional datasets. To create a more holistic approach, the automated processing of social media images should be used in conjunction with other biodiversity monitoring methods and datasets. While BioCLIP appears to offer the best solution to reducing false negative sightings in noisy image datasets from social media, researchers must consider the specific context and requirements of their studies when selecting whether a foundation vision model is suitable for their research, and where possible, complement the use of these models with manual validation or additional dataset filtering to enhance accuracy.

References

Allain, S. (2019). Mining Flickr: A method for expanding the known distribution of invasive species.

Herpetological Bulletin, *148, Summer 2019*, 11–14. https://doi.org/10.33256/hb148.1114

August, T. A., Pescott, O. L., Joly, A., & Bonnet, P. (2020). AI Naturalists Might Hold the Key to Unlocking Biodiversity Data in Social Media Imagery. *Patterns*, *1*(7). https://doi.org/10.1016/j.patter.2020.100116

Barve, V. (2014). Discovering and developing primary biodiversity data from social networking sites: A novel approach. *Ecological Informatics*, *24*, 194–199. https://doi.org/10.1016/j.ecoinf.2014.08.008

Bradley, J., Lapp, H., & Campolongo, E. G. (2024). *Pybioclip* (Version 1.0.0) [Python]. https://doi.org/10.5281/zenodo.13151194

Campbell, C. J., Barve, V., Belitz, M. W., Doby, J. R., White, E., Seltzer, C., Di Cecco, G., Hurlbert, A. H., & Guralnick, R. (2023). Identifying the identifiers: How iNaturalist facilitates collaborative, research-relevant data generation and why it matters for biodiversity science. *BioScience*, *73*(7), 533–541. https://doi.org/10.1093/biosci/biad051

Cardoso, A. S., Malta-Pinto, E., Tabik, S., August, T., Roy, H. E., Correia, R., Vicente, J. R., & Vaz, A. S. (2024). Can citizen science and social media images support the detection of new invasion sites? A deep learning test case with *Cortaderia selloana*. *Ecological Informatics*, *81*, 102602. https://doi.org/10.1016/j.ecoinf.2024.102602

Di Cecco, G. J., Barve, V., Belitz, M. W., Stucky, B. J., Guralnick, R. P., & Hurlbert, A. H. (2021). Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. *BioScience*, *71*(11), 1179–1188. https://doi.org/10.1093/biosci/biab093

Edwards, T., Jones, C. B., Perkins, S. E., & Corcoran, P. (2021). Passive citizen science: The role of social media in wildlife observations. *PLOS ONE*, *16*(8), e0255416. https://doi.org/10.1371/journal.pone.0255416

Fernandes, A. F. A., Dórea, J. R. R., & Rosa, G. J. de M. (2020). Image Analysis and Computer Vision Applications in Animal Sciences: An Overview. *Frontiers in Veterinary Science*, *7*. https://doi.org/10.3389/fvets.2020.551269

Fox, N., August, T., Mancini, F., Parks, K. E., Eigenbrod, F., Bullock, J. M., Sutter, L., & Graham, L. J. (2020). “photosearcher” package in R: An accessible and reproducible method for harvesting large datasets from Flickr. *SoftwareX*, *12*, 100624. https://doi.org/10.1016/j.softx.2020.100624

Fox, N., Di Minin, E., Carter, N., Tomkins, S., & Van Berkel, D. (2024). Artificial Intelligence and Crowdsourced Social Media Data for Biodiversity Monitoring and Conservation. In A. Olanrewaju & S. Bruno (Eds.), *Advancements in Architectural, Engineering, and Construction Research and Practice* (pp. 43–50). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-59329-1_4

Fox, N., Graham, L. J., Eigenbrod, F., Bullock, J. M., & Parks, K. E. (2021). Enriching social media data allows a more robust representation of cultural ecosystem services. *Ecosystem Services*, *50*, 101328. https://doi.org/10.1016/j.ecoser.2021.101328

Fox, N., & Van Berkel, D. (2024). Identifying invasive species sightings from GeoAI-validated social media posts. *I-GUIDE Forum 2024: Convergence Science and Geospatial AI for Environmental Sustainability*. I-GUIDE Forum. https://doi.org/10.5703/1288284317801

Ghermandi, A., Langemeyer, J., Van Berkel, D., Calcagni, F., Depietri, Y., Egarter Vigl, L., Fox, N., Havinga, I., Jäger, H., Kaiser, N., Karasov, O., McPhearson, T., Podschun, S., Ruiz-Frau, A., Sinclair, M., Venohr, M., & Wood, S. A. (2023). Social media data for environmental sustainability: A critical review of opportunities, threats, and ethical use. *One Earth*, *6*(3), 236–250. https://doi.org/10.1016/j.oneear.2023.02.008

Groom, Q. J., & Whild, S. J. (2017). Characterisation of false-positive observations in botanical surveys. *PeerJ*, *5*, e3324. https://doi.org/10.7717/peerj.3324

Hartmann, M. C., Schott, M., Dsouza, A., Metz, Y., Volpi, M., & Purves, R. S. (2022). A text and image analysis workflow using citizen science data to extract relevant social media records: Combining red kite observations from Flickr, eBird and iNaturalist. *Ecological Informatics*, *71*, 101782. https://doi.org/10.1016/j.ecoinf.2022.101782

Jeawak, S., Jones, C., & Schockaert, S. (2018). *Mapping wildlife species distribution with social media: Augmenting text classification with species names* (S. Winter, A. Griffin, & M. Sester, Eds.; Vol. 114, p. 34:1-34:6). Schloss Dagstuhl/Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.GISCIENCE.2018.34

Johnston, A., Matechou, E., & Dennis, E. B. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *14*(1), 103–116. https://doi.org/10.1111/2041-210X.13834

Kühl, H. S., Bowler, D. E., Bösch, L., Bruelheide, H., Dauber, J., Eichenberg, D., Eisenhauer, N., Fernández, N., Guerra, C. A., Henle, K., Herbinger, I., Isaac, N. J. B., Jansen, F., König-Ries, B., Kühn, I., Nilsen, E. B., Pe'er, G., Richter, A., Schulte, R., … Bonn, A. (2020). Effective Biodiversity Monitoring Needs a Culture of Integration. *One Earth*, *3*(4), 462–474. https://doi.org/10.1016/j.oneear.2020.09.010

Kulkarni, R., & Di Minin, E. (2023). Towards automatic detection of wildlife trade using machine vision models. *Biological Conservation*, *279*, 109924. https://doi.org/10.1016/j.biocon.2023.109924

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., & Gao, J. (2024). Multimodal Foundation Models: From Specialists to General-Purpose Assistants. *Foundations and Trends® in Computer Graphics and Vision*, *16*(1–2), 1–214. https://doi.org/10.1561/0600000110

Mancini, F., Coghill, G. M., & Lusseau, D. (2019). Quantifying wildlife watchers' preferences to investigate the overlap between recreational and conservation value of natural areas. *Journal of Applied Ecology*, *56*(2), 387–397. https://doi.org/10.1111/1365-2664.13274

Mangachena, J. R., & Pickering, C. M. (2023). Why are some animals popular with wildlife tourists: Insights from South Africa. *Journal of Ecotourism*, *22*(2), 312–328. https://doi.org/10.1080/14724049.2021.2019261

Marshall, B. M., & Strine, C. T. (2019). Exploring snake occurrence records: Spatial biases and marginal gains from accessible social media. *PeerJ*, *7*, e8059. https://doi.org/10.7717/peerj.8059

OpenStreetMap. (2024). *OpenStreetMap contributors. Available under the Open Database Licence from: Openstreetmap.org. Data mining by Overpass turbo. Available at overpass-turbo.eu.* [Dataset].

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision* (arXiv:2103.00020). arXiv. https://doi.org/10.48550/arXiv.2103.00020

Sbragaglia, V., Espasandín, L., Coco, S., Felici, A., Correia, R. A., Coll, M., & Arlinghaus, R. (2022). Recreational angling and spearfishing on social media: Insights on harvesting patterns, social engagement and sentiments related to the distributional range shift of a marine invasive species. *Reviews in Fish Biology and Fisheries*, *32*(2), 687–700. https://doi.org/10.1007/s11160-022-09699-7

Schirpke, U., Ghermandi, A., Sinclair, M., Van Berkel, D., Fox, N., Vargas, L., & Willemen, L. (2023). Emerging technologies for assessing ecosystem services: A synthesis of opportunities and challenges. *Ecosystem Services*, *63*, 101558. https://doi.org/10.1016/j.ecoser.2023.101558

Schirpke, U., Tasser, E., Ebner, M., & Tappeiner, U. (2021). What can geotagged photographs tell us about cultural ecosystem services of lakes? *Ecosystem Services*, *51*, 101354. https://doi.org/10.1016/j.ecoser.2021.101354

Schmeller, D. S., Böhm, M., Arvanitidis, C., Barber-Meyer, S., Brummitt, N., Chandler, M., Chatzinikolaou, E., Costello, M. J., Ding, H., García-Moreno, J., Gill, M., Haase, P., Jones, M., Juillard, R., Magnusson, W. E., Martin, C. S., McGeoch, M., Mihoub, J.-B., Pettorelli, N., … Belnap, J. (2017). Building capacity in biodiversity monitoring at the global scale. *Biodiversity and Conservation*, *26*(12), 2765–2790. https://doi.org/10.1007/s10531-017-1388-7

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., & Keutzer, K. (2021). *How Much Can CLIP Benefit Vision-and-Language Tasks?* (arXiv:2107.06383). arXiv. http://arxiv.org/abs/2107.06383

Soriano-Redondo, A., Correia, R. A., Barve, V., Brooks, T. M., Butchart, S. H. M., Jarić, I., Kulkarni, R., Ladle, R. J., Vaz, A. S., & Minin, E. D. (2024). Harnessing online digital data in biodiversity monitoring. *PLOS Biology*, *22*(2), e3002497. https://doi.org/10.1371/journal.pbio.3002497

Spooner, S. L., & Stride, J. R. (2021). Animal-human two-shot images: Their out-of-context interpretation and the implications for zoo and conservation settings. *Zoo Biology*, *40*(6), 563–574. https://doi.org/10.1002/zoo.21636

Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., Dong, L., Dahdul, W. M., Stewart, C., Berger-Wolf, T., Chao, W.-L., & Su, Y. (2024). *BioCLIP: A Vision Foundation Model for the Tree of Life*. 19412–19424. https://openaccess.thecvf.com/content/CVPR2024/html/Stevens_BioCLIP_A_Vision_Foundation_Model_for_the_Tree_of_Life_CVPR_2024_paper.html

Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports*, *7*(1), 17615. https://doi.org/10.1038/s41598-017-18007-4

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, *233*, 298–315. https://doi.org/10.1016/j.biocon.2019.01.023

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., van Horn, G., Crofoot, M. C., Stewart, C. V., & Berger-Wolf, T. (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications*, *13*(1), 792. https://doi.org/10.1038/s41467-022-27980-y

Wäldchen, J., & Mäder, P. (2018). Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review. *Archives of Computational Methods in Engineering*, *25*(2), 507–543. https://doi.org/10.1007/s11831-016-9206-z

Wang, J., Chan, K. C. K., & Loy, C. C. (2023). Exploring CLIP for Assessing the Look and Feel of Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(2), Article 2. https://doi.org/10.1609/aaai.v37i2.25353

Weinstein, B. G. (2018). A computer vision for animal ecology. *Journal of Animal Ecology*, *87*(3), 533–545. https://doi.org/10.1111/1365-2656.12780

Wiens, J. J. (2023). How many species are there on Earth? Progress and problems. *PLOS Biology*, *21*(11), e3002388. https://doi.org/10.1371/journal.pbio.3002388

Wolf, T., Debut, L., Sanh, V., Chaumond, J., & Delangue, C. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38--45). Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.emnlp-demos.6